Supplementary information, Data S2 False positive (FP) and false negative (FN) rates for variant calling

In our previous study, we sequenced a collection of 614 segments from six individuals using traditional Sanger sequencing[38]. Sequencing all segments from all individuals would be quite challenging. We thus used this verification data to infer/project the overall false positive (FP)/negative (FN) rates for the variant calling.

In order to predict the trend of FP/FN as a function of the number of individuals, we randomly subsampled 1-6 samples and computed the overall FP/FN rate for variant calling. The Sanger sequencing result was used as the benchmark set for the FP and FN calculation. From Supplementary information, Figure S2, we can see that, we are missing around 10% of the variants seen in the Sanger sequencing data (i.e., FN rate is about 10%). Within the total missed SNPs, the vast majority of the false negatives are due to singleton variants segregating in the sample. The amount of FP is quite limited and is about 3-4%. This overall FN/FP matches the general performances observed in the GATK package and the 1000 genome project[3, 4].